유전자 알고리즘을 이용한 KeyGraph 알고리즘의 데이터 분할

이영설[O] 조성배
연세대학교 컴퓨터과학과
tiras[O]@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Sentence segmentation of KeyGraph using genetic algorithm

Youngseol Lee[0], Sung-Bae Cho
Department of Computer Science, Yonsei University

요 약

키그래프는 데이터 패턴 속에서 인간의 의사결정이나 미래에 닥쳐올 변화에 영향을 주지만 자주 발생하지 않는 희소성이 있는 사건을 발견하기 위한 알고리즘이다. 키그래프는 지진예측, 논문요약, 파일탐색, 그리고 중요한 URL 추출 등에 이용되었다. 데이터 분할을 통한 클러스터의 형성은 키그래프의 성능에 가장 큰 영향을 끼치는 요소 중의 하나이다. 본 논문에서는 유전자 알고리즘을 이용하여 키그래프의 성능을 향상시킬 수 있는 최적의 데이터 분할을 찾아내는 방법을 제안한다. 제안한 방법의 가능성을 보여주기 위하여 모바일 기기 사용자로부터 수집한 방문 장소 데이터에 제안하는 방법을 적용하여 키그래프의 성능이 향상되는 것을 보인다.

1. 서 론

Chance discovery 는 인간의 의사결정이나 미래에 닥쳐올 변화에 영향을 주지만 자주 발생하지 않는 희소성이 있는 사건을 발견하기 위한 연구이다[1]. Chance는 발생 빈도가 낮고 인식하기 어렵지만 미래의 변화에 중요한 역할을 하게 될 사건이다. Chance는 인간의 의사결정에 영향을 끼치며, 긍정적인 기회로서 작용하거나 부정적인 위기가 된다. 키그래프는 데이터 패턴 속에서 이런 Chance에 해당하는 데이터를 찾아내기 위한 알고리즘이다. 키그래프는 지진예측, 논문요약, 파일탐색, 그리고 중요한 URL 추출 등에 이용되어 왔다. 키그래프 알고리즘이 Chance에 해당하는 데이터를 잘 찾아내기 위해서는 먼저 키그래프를 적용하기 위한 데이터가 주의 깊게 선택되어야 하며 데이터에 발생하는 변화를 기준으로 데이터를 분할함으로써 관계가 있는 데이터들끼리 클러스터를 형성하도록 해야 한다.

본 논문에서는 유전자 알고리즘을 바탕으로 데이터를 분할하여 키그래프가 Chance를 발견하는 능력을 향상시키는 방법을 제안한다.

1.1. 키그래프

키그래프(KeyGraph)는 개념적으로 두 단계의 과정을 거쳐 Chance를 발견한다[2].

(1) 클러스터 형성 단계

첫 번째 단계는 관계가 있는 데이터를 묶어 클러스터를 형성하는 것이다. 데이터 분할은 데이터에 나타나는 큰 변화를 기준으로 전체 데이터를 하위 데이터 집합으로 분할하는 과정이다. 하위 데이터 집합 안에 존재하는 데이터들은 서로 관계가 있는 것으로 간주되며, 하위 데이터 집합 안에 자주 함께 나타나는 데이터들일수록 그 사이의 관계가 강해지고 클러스터를 형성하게 된다. 그림 1은 데이터의 분할과 클러스터 생성 사이의 관계를 간단히 표현하고 있다.

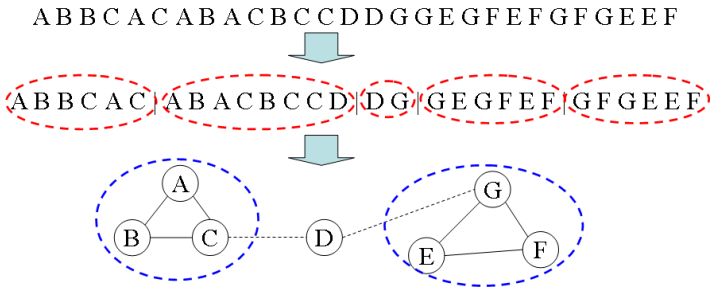

그림 1. 데이터 분할과 클러스터 형성

(2) Chance의 후보 발견 단계

두 번째 단계는 Chance의 후보가 되는 데이터를 발견하는 것이다. 키그래프 알고리즘은 하나의 클러스터 내부에 존재하는 데이터들은 그 데이터를 발생시키는 공통적인 원인이 있을 것이라고 추정한다. 하나의 클러스터에서 집중적으로 나타나지 않고, 여러 클러스터에 걸쳐서 공통적으로 나타나는 데이터는 클러스터와 클러스터를 연결하는 연결점이며, 공통된 원인으로부터 비롯된 하나의 데이터 집합이 다른 데이터 집합으로 변화하는 분기점이 된다고 볼 수 있다. 이런 데이터가 발견되면 전체 데이터에 변화가 발생할 가능성이 있으며 따라서 이런 데이터는 Chance의 후보로 간주된다. 그림 2는 여러 클러스터와 관계가 있는 데이터를 개념적으로 나타내고 있다.

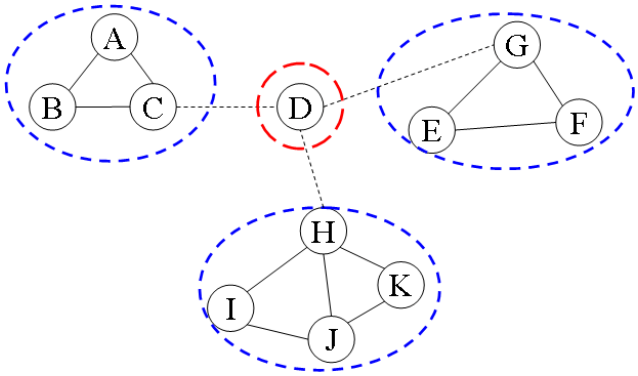

그림 2. 키그래프에서 Chance 의 발견

1.2 키그래프의 성능에 영향을 끼치는 요소

(1) 클러스터 형성 단계

이 단계에서 키그래프가 Chance를 발견하는 능력에 영향을 끼치는 요소는 데이터 전처리 과정과 데이터 분할 과정이다. 데이터 전처리 과정이 잘 진행되지 못하면 실제로는 거의 의미 없는 데이터가 데이터 집합에 다수 포함될 수 있으며 키그래프 알고리즘이 의미 없는 데이터를 Chance의 후보로 선택할 가능성이 높아지게 된다.

만약 데이터 분할이 잘 이루어지지 못한다면 관계있는 데이터가 관계없는 데이터들과 하나의 클러스터로 묶이게 되고 데이터로부터 Chance를 발견할 가능성이 낮아지게 된다.

(2) Chance의 후보 발견 단계

이 단계에서 키그래프의 결과에 영향을 끼치는 요소는 데이터 분할 과정에서 분할된 하위 데이터 집합에 자주 나타나는 데이터의 개수와 여러 클러스터와 관계있는 데이터의 개수이다. 이 숫자를 조정하는 것으로 키그래프의 결과를 조정할 수 있으며 이미 이 숫자들을 진화 알고리즘으로 조정하여 사용자가 더 이해하기 쉬운 그래프 구조를 생성하는 연구가 이루어 졌다[3].

1.3 키그래프의 데이터 분할

이전의 연구에서 데이터를 분할하는 기준은 데이터 도메인에 따라서 달라졌다. 데이터가 특정 기준으로 구분되는 데이터일 경우에는 데이터를 구분하는 기준을 그대로 데이터 분할에 이용하였다. 예를 들어 논문으로부터 논문의 핵심 키워드를 추출하는 연구에서는 논문에 나타나는 단어를 데이터로 치환하고, 단어의 집합을 구분하기 위하여 문장 단위로 데이터를 분할하였다. 하나의 문장은 한 가지 의도를 표현하기 위한 단어의 집합이므로 한 문장마다 데이터를 분할하는 것은 합리적인 방법이다[4]. 특정한 이슈와 관련된 웹페이지들에서 중요한 URL을 추출하기 위한 연구에서는 각 URL 주소를 데이터로 간주하였으며, 웹페이지 단위로 데이터를 분할하였다. 여기서는 하나의 웹페이지가 한 가지 주제를 표현하기 위해서 만들어진 것으로 추정되므로 웹페이지 마다 데이터를 분할하는 것도 합리적인 방법이다[5].

그러나 이렇게 데이터를 구분할 수 있는 기준이 없을 경우에는 데이터 분할을 위하여 추가적인 정보를 사용하거나 전문가의 의견에 따라서 데이터를 분할하기 위한 기준을 정하였다. 예를 들어 지진데이터로부터 중요한 진앙을 찾기 위한 연구에서는 커다란 지진이나 지각 변동을 기준으로 데이터를 분할하였으며,[2] PC 사용자의 파일 사용 기록으로부터 사용자에게 중요한 파일을 추출하는 연구에서는 사용자의 로그인, 로그아웃 시점을 기준으로 사용자의 파일 사용 기록을 분할하였다[6]. 그러나 데이터를 분할하기 위해 획득한 추가적인 정보가 항상 최적의 데이터 분할을 제공하지는 않으며, 전문가가 정한 데이터 분할 기준을 사용할 경우 전문가가 데이터에 나타나는 변화 가운데 어떤 것을 분할의 기준으로 간주하느냐에 따라서 키그래프가 Chance를 찾아내지 못할 수도 있다. 따라서 본 논문에서는 진화 알고리즘을 이용하여 Chance를 발견하기 위한 최적의 데이터 분할을 찾는 방법을 제안한다.

2. 제안하는 방법

본 논문에서 제안하는 방법은 진화 알고리즘을 이용하여 키그래프가 Chance를 발견하기 위한 최적의 데이터 분할을 찾는 것이다.

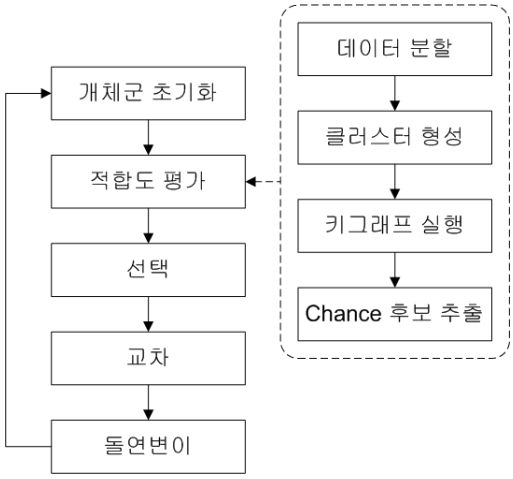

그림 3. 전체 시스템 구성도

그림 3은 전체 시스템의 구성을 간략히 나타내고 있다. 여기서 데이터 분할은 유전자 알고리즘의 유전자로 표현되고 데이터 분할의 집합이 개체군이 된다. 적합도가 높은 유전자를 찾기 위해서 유전자가 의미하는 데이터 분할을 이용하여 키그래프를 실행하고 실행 결과의 적합도를 평가한다. 최종적으로 적합도가 가장 높은 유전자를 선택하여 선택된 유전자가 실제로 키그래프가 Chance를 발견하는 능력을 향상시키는지 평가한다.

여기서 유전자 알고리즘을 사용하는 이유는 그것이 방대한 탐색 공간에서 최적의 해를 찾기에 적합한 방법이기 때문이다. 만약 100개의 데이터가 존재한다면 이 데이터를 분할하는 방법의 수는 총 2^100 개이며, 만약 1000개의 데이터가 있다면 이 데이터를 분할하는 방법의 수는 2^1000 개가 있게 된다. 이런 거대한 탐색 공간에서 원하는 해를 얻기 위해서 유전자 알고리즘은 적합한 방법이다.

2.1 초기 개체군의 결정

진화 알고리즘은 방대한 탐색 공간을 가지는 문제에 적합한 방법이다. 그러나 2^1000 은 지나치게 큰 탐색 공간이다. 이런 탐색 공간에서 좋은 해를 찾아내기 위해서는 초기 개체군의 수를 상당히 크게 하거나 오랜 시간 동안 돌연변이를 발생시키야 한다. 그러나 개체군의 수가 커질수록 진화에 걸리는 시간과 메모리 사용량이 늘어나서 시간적으로나 물리적으로 한계에 봉착하게 된다.

반면 개체군 숫자를 너무 적게 주게 되면 랜덤하게 개체군을 구성하였을 경우, 원하는 해에서 완전히 동떨어진 결과를 얻게 될 가능성이 있다. 결국, 비교적 적은 수의 초기 개체군으로 원하는 결과를 얻기 위해서는 초기 개체군을 랜덤하게 설정하기보다는 휴리스틱을 이용해서 합리적인 초기 개체군을 설정하여 진화를 시작하는 것이 효과적인 방법이다.

따라서 여기서는 개체군을 구성하기 위하여 이전 연구에서 추출한 랜드마크를 이용하여 기본 데이터 분할을 만들었다[7]. 랜드마크는 사용자가 스마트 폰을 사용하면서 수집된 각종 데이터(통화기록, 문자메시지 기록, 사진찍은 정보, GPS 기록 등등)를 바탕으로 추론된 사용자의 행동이나 감정이다. 랜드마크를 기준으로 분할된 데이터 분할에서 단 하나의 유전자 코드만을 변경하여 초기 개체군을 생성하였다.

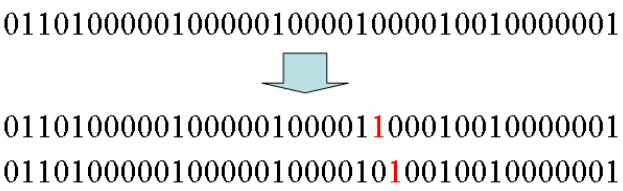

그림 4. 초기 개체군 생성

2.2 데이터 분할의 적합도 평가 기준 결정

일반적으로 더 좋은 데이터 분할은 적절한 클러스터를 생성하게 만들어 키그래프가 Chance에 해당하는 데이터를 잘 찾을 수 있도록 한다. 데이터 분할의 적합도 평가는 본래 데이터 분할이 키그래프가 클러스터의 구조를 잘 표현하고, Chance를 발견할 수 있도록 하는지를 기준으로 평가해야 한다.

키그래프와 진화알고리즘을 사용한 이전의 연구에서는 적합도를 평가하기 위해서 대화형 사용자 인터페이스를 사용하여 진화를 통해서 나온 결과를 사람이 직접 평가하였다.[7] 그리고 사람이 평가한 값을 기준으로 키그래프를 진화시켰다. 그러나 진화의 결과를 사람이 일일이 평가하는 방법은 만약 진화의 결과가 육안으로 판단하기 어려울 정도의 차이밖에 없을 경우 정확한 평가가 불가능해진다. 여기서는 두 가지 기준으로 유전자의 적합도를 평가하고 진화를 수행하였다.

(1) Chance의 후보의 개수로 적합도 평가

키그래프에서 좋은 결과는 Chance에 해당하는 데이터를 많이 발견하는 것이다. 여기서는 Chance의 후보를 많이 추출할수록 Chance에 해당하는 데이터를 많이 얻을 가능성이 증가한다고 가정하고 Chance의 후보로 추출된 데이터의 수로 유전자의 적합도를 평가한다.

(2) Column 값으로 적합도 평가

Column 값은 하나의 클러스터를 대표하는 데이터와 발생 빈도가 낮으면서 여러 클러스터와 관련된 데이터의 관계를 수치화한 값으로 이 값이 큰 데이터 일수록 Chance일 가능성이 커진다. 따라서 키그래프의 실행 결과로 나온 Column 값의 총합이 클수록 유전자의 적합도가 높다고 가정할 수 있다. 다음 공식 (1)은 Column 값을 계산하는 방법을 간략하게 나타내고 있다.

$$column(w_i, w_j) = \sum_{s \in D} \min(|w_i|, |w_j|) \quad (1)$$

S : 데이터 분할에 의해 분할된 데이터 집합

D : 전체 데이터 집합

w_i : 데이터 클러스터를 대표하는 데이터의 빈도

w_j : 여러 클러스터와 관련된 데이터의 빈도

3. 실험 및 결과

3.1 실험 환경

본 논문에서는 제안하는 방법의 가능성을 보여주기 위해서 스마트폰에서 수집한 방문 장소 데이터를 대상으로 실험을 진행하여 Chance에 해당하는 장소를 추출하는 것을 보인다. 프로그램은 VC++ 6.0 으로 제작되었고, 실험 환경은 윈도우 2003 서버에서 진행되었다. 실험에 이용된 데이터는 스마트 폰 사용자에게 수집된 방문 장소 목록이 이용되었으며, 최종적인 평가를 위해 Subjective Test를 진행하였다. 실험에 이용된 데이터에서 사용자가 방문한 장소 데이터의 총 개수는 828개이며 장소의 종류는 모두 57개이다.

3.2 실험 방법

실험에서는 총 828개의 방문 장소 목록을 분할하는 데이터 분할의 초기 개체군을 랜드마크를 이용하여 생성하고, 만들어진 개체군에 유전자 알고리즘을 적용하여 적합도가 높은 유전자를 추출하였다. 또 최종적으로 선택된 유전자가 키그래프의 성능을 향상시키는지 평가하기 위하여 실험 참가자가 방문한 장소의 Unusualness를 평가하고 적합도가 높은 데이터 분할을 키그래프에 적용할 경우 Chance에 해당하는 장소를 발견하는 성능이 향상되는지 살펴보았다. 실험에 사용된 유전자 알고리즘의 교차율은 0.7, 돌연변이율은 0.01로 하였다.

3.3 실험 결과 - 적합도 평가

(1) Chance의 후보 개수로 적합도 평가

Chance의 후보가 되는 데이터의 개수로 데이터 분할의 적합도를 평가하여 진화시킬 경우 발생하는 문제점은 키그래프를 통해서 얻는 Chance의 후보 개수가 기본적으로 큰 차이가 없다는 점이다. 따라서 Chance의 후보 개수만 가지고 평가할 경

우 진화가 잘 이루어지지 못하였다. 여기서는 유전자들 사이에 미세한 차이를 만들어 주기 위하여 Key 값을 추가적으로 사용하였다. Key 값은 특정 데이터가 Chance의 후보가 될 가능성을 나타내는 확률 값으로 이 값이 높을수록 Chance를 잘 발견할 수 있다고 판단하였다. 공식 (2), (3), (4) 는 Key 값을 간략히 나타내고 있다. 적합도 평가는 (Chance의 후보 개수) + (Key 값의 총합) 으로 평가하였다.

$$based(w,g) = \sum_{s \in D} |w|_s |g - w|_s \quad (2)$$

$$neighbors(g) = \sum_{s \in D} \sum_{w \in s} |w|_s |g - w|_s \quad (3)$$

$$key(w) = 1 - \prod_{g \subset G} (1 - \frac{based(w,g)}{neighbors(g)}) \quad (4)$$

그림 5는 세대가 지남에 따라 변화하는 적합도 값을 나타내고 있다. 그림 5에서 굵은 실선은 최대 적합도를 나타내며, 가는 실선은 평균 적합도를 나타낸다.

(2) Column 값으로 적합도 평가

Column 값은 클러스터와 데이터 사이의 관계를 나타내는 값으로 이 값이 클수록 데이터가 Chance로 선택될 가능성이 높아진다. 그림 6은 세대에 따른 적합도의 변화를 보여준다. 굵은 실선은 세대별 최대 적합도를 나타내며 가는 실선은 세대별 평균 적합도를 나타낸다.

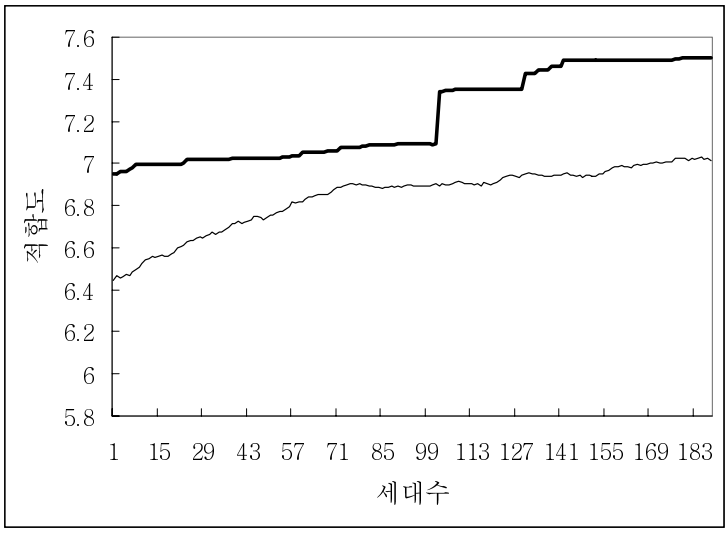

그림 5. 적합도의 변화

3.4 실험 결과 - 최종 평가

(1) 최종 평가를 위한 Subjective Test

최종적인 평가를 위해서는 적합도가 높은 유전자가 실제로 Chance 에 해당하는 발견하는지 확인해야 한다. Chance에 해당하는 장소는 평소에 자주 방문하지 않으면서 일단 방문하면 사용자의 행동이나 감정에 변화를 유발시키는 장소이다. 이 실험에서는 방문한 장소에서 사용자의 상태에 변화를 유발시킬 수 있는 비일상적인 행동을 평가하기 위하여 Subjective Test를 진행하였다. 표 1은 방문 장소에 따른 방문 횟수와 비일상적인 정도를 정리한 것이다. (비일상도 / 방문횟수)를 계산하여 이 값이 클수록 Chance에 가까운 장소로 판단할 수 있다.

$$unusual(p) = rarity(p) / frequency(p) \quad (5)$$

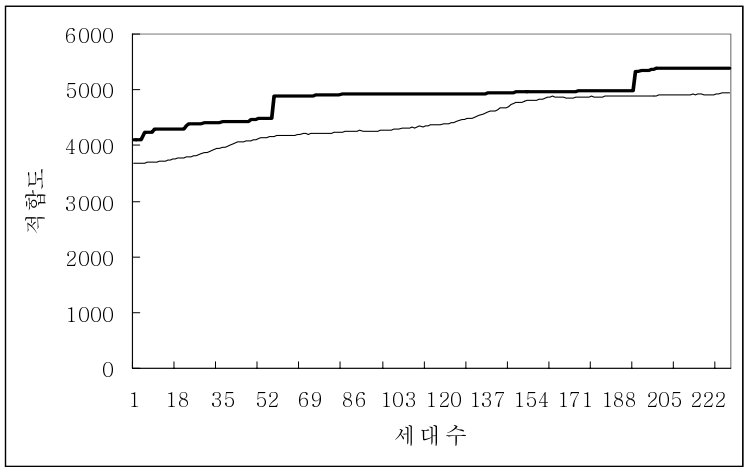

그림 6. 적합도 변화

표 1. 장소에 따른 비일상도와 방문횟수

순번	장소명	방문 횟수	비일상도	비일상도/ 방문횟수
1	square	90	4.85	0.05
2	women university	58	2.83	0.05
3	park	35	4.1	0.12
4	home	32	2.37	0.07
5	Cultural Street	23	3.94	0.17
6	student union building	16	5.25	0.33
7	student central military school	15	5.33	0.36
8	college of engineering	12	2.96	0.25
9	library	10	3.17	0.32
10	college of music	8	4.5	0.56
11	chinese restaurant	7	3.5	0.5
12	middle school	7	4.3	0.61
13	store	6	3.87	0.65
14	Yonsei engineering research center	6	2.83	0.47
15	coffee shop	4	3.72	0.93
16	department store	4	5	1.25
17	fashion shopping Mall	4	5	1.25
18	general classroom building	4	2.6	0.65
19	Trade Center	3	4.17	1.39
20	electronic market	2	4.75	2.38

(2) Chance의 후보 개수로 적합도 평가

그림 8은 세대가 지남에 따라서 Unusualness 값이 변화하는 상태를 나타내고 있다. Chance의 후보 개수로 적합도를 평가하여 진화시킨 경우, Unusualness 값이 한 번 크게 증가한 이후

에는 다시 증가하는 모습을 보여주지 못하고 있다. 결국 적합도를 Chance의 후보의 개수로 평가하여 진화시키는 방법은 좋은 방법이 아니라는 것을 알 수 있다. 특히 Chance의 후보의 개수가 그다지 변화하지 않는 경우 유전자 알고리즘의 효용성은 매우 떨어지게 된다.

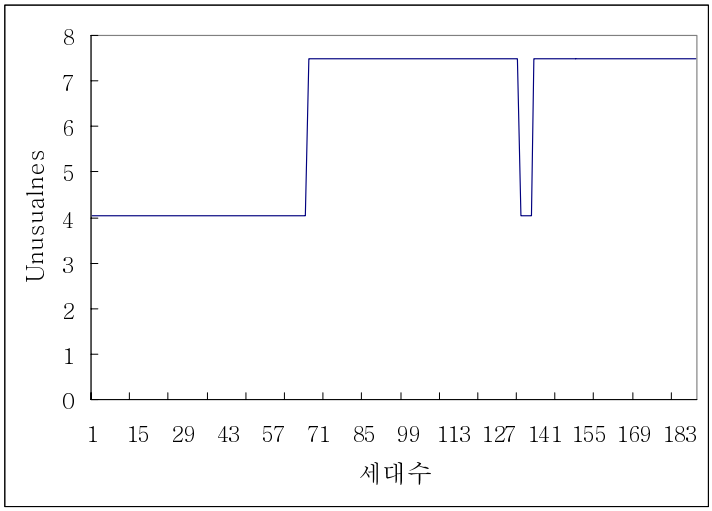

그림 7. Unusualness의 변화

(3) Column 값으로 적합도 평가

그림 9는 세대가 지남에 따라서 Unusualness 값이 변화하는 상태를 보여주고 있다. 여기서는 Unusualness 값이 계단식으로 증가하는 모습을 보여준다. 그러나 일정 이상 시간이 흐르면 더 이상 증가하지 않고 수렴하는 형태를 보여 준다.

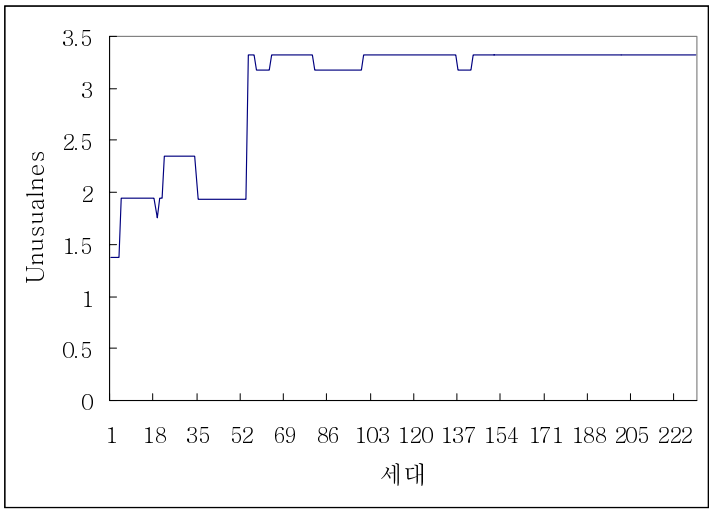

그림 8. Unusualness의 변화

4. 결론 및 향후 연구

본 논문에서는 키그래프 알고리즘에 유전자 알고리즘을 적용하여 키그래프 알고리즘의 성능을 향상시키려고 시도하였다. 제안하는 방법의 가능성을 보이기 위하여 모바일 기기의 사용자가 방문한 장소 데이터를 대상으로 유전자 알고리즘을 적용하여 진화를 수행하기 전보다 진화한 이후에 키그래프의 성능이 향상됨을 보였다. 결론적으로 Chance의 후보를 많이 추출하거나 Column 값을 증가시키는 방향으로 진화하도록 하여 키그래프의 성능을 일정 수준 향상시킬 수 있었다. 그러나 이 두 가지 기준이 항상 키그래프의 성능을 향상시키는 것은 아니며, 이러한 유전자 알고리즘을 진행하기에 앞서 먼저 데이터 전처리가 세심하게 진행되어야 한다. 또한 유전자 알고리즘을 위한 초기 개체군을 생성할 때 납득할 수 있는 기준으로 초기 개체군이 선정되어야 한다. 그렇지 않으면, 넓은 탐색 공간에서 원하는 해를 구할 수 없게 된다. 향후 연구로는 앞서 제시한 기준 이외에 키그래프의 성능을 판단할 수 있는 객관적인 지표가 있는 지 확인하고 그것을 통하여 진화 시키는 것과 몇 가지 기준을 동시에 적용하여 진화시킴으로서 더 나은 성능을 얻을 수 있는지 확인하는 것이 있다.

참고 문헌

[1] K.-J. Kim, S.-B. Cho, "Uncertainty reasoning and chance discovery," *Chance Discovery in Real World Decision Making (Studies in Computational Intelligence 30)*, pp. 85-102, Springer, 2006, 10.

[2] Y. Ohsawa, "Keygraph as risk explorer in earthquake-sequence," *Journal of Contingencies and Crisis Management*, vol. 10, pp. 119-128, 2002.

[3] X. Llor`a, N. Matsumura, E. D. Goldberg, Y. Ohsawa, K. Ohnishi, and A. Gonzales, "Discovering Chance Scenarios using Small-World KeyGraphs and Evolutionary Computation," *First European Workshop on Chance Discovery (EWCD'2004), ECAI press*, pp 51 - 61, 2004.

[4] Y. Ohsawa, N. E. Benson and M. Yachida, "Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor," *Proc. Advanced Digital Library Conference (IEEE ADL'98)*, pp. 12-18, 1998.

[5] N. Matsumura, Y. Matsuo, Y. Ohsawa, and M. Ishizuka, "Discovering Emerging Topics from WWW," *Journal of Contingencies and Crisis Management*, Vol. 10, pp. 73-81, 2002.

[6] Y. Ohsawa, "Get timely files from visualized structure of your working history," *Knowledge-Based Intelligent Information Engineering Systems*, pp. 546-549, 1999.

[7] K.-S. Hwang and S.-B. Cho, "Modular Bayesian Networks for Inferring Landmarks on Mobile Daily Life," *The 19th Australian Joint Conference on Artificial Intelligence*, pp. 929-933, 2006.

[8] N. Okazaki and Y. Ohsawa, "Polaris: An Integrated Data Miner for Chance Discovery," *In Proceedings of The Third International Workshop on Chance Discovery and Its Management*, Crete, Greece 2003.